

# Automatic Enrichment of Stream Networks with Primary Paths for Use in the United States National Atlas

Barbara P. Battenfield\*, Lawrence V. Stanislawski\*\*, Christopher Anderson-Tarver\* and Michael J. Gleason\*

\* Department of Geography, University of Colorado, Boulder Colorado USA

\*\* Center for Excellence in Geospatial Information Science, United States Geological Survey, Rolla Missouri USA

**Abstract.** This paper presents a process of enrichment and generalization for automatic demarcation of primary paths through a complex stream network, in support of data production for the U.S. National Atlas. The primary path delineation forms the basis for a reduced scale version of hydrography appropriate for base and thematic mapping at smaller scales. Advantages of producing 1:1million scale National Atlas data from 1:24,000 source data are to provide a higher source of accuracy for the National Atlas data, to improve data currency, and to establish feature-level linkages between source and target scale databases, in preparation for transitioning national hydrography datasets into a multiple representation database (MRDB) framework.

**Keywords:** generalization, hydrography, USGS National Atlas, MRDB

## 1. Introduction

Hydrography comprises a commonly included vector layer in topographic base mapping. Because it is highly sensitive to scale change, hydrographic data generalization has emerged as a primary focus among cartographers, hydrologists, and agencies that produce geospatial data for use at multiple scales. Challenges encountered when generalizing hydrography include preservation of flow direction, logical reflection of channel hierarchy, horizontal integration (maintaining connectivity among channels and polygonal waterbodies), and vertical integration with other data layers, for example terrain contours and transportation networks. This paper reports a genera-

lization experiment for a hydrographic network that relates to preserving channel hierarchy.

To minimize data maintenance and integration, the United States Geological Survey (USGS) is working to automate generalization of its most detailed datasets to smaller scales to support multi-scale display and delivery of these data. This includes generalization of the high-resolution (HR) layer of the National Hydrography Dataset (NHD), which is compiled at 1:24,000 (24K) or larger scales (1:63,360 in Alaska). The HR NHD has been undergoing updates over the past several years to improve the data and add detail in areas. However, many areas are covered with legacy data compiled over numerous years with differing conditions. Consequently, the HR NHD layer is a multi-scale dataset with compilation variations that are evident in areas.

The HR NHD data does not include a specific attribute that identifies primary stream paths through the flow network. Primary paths are cartographically important at larger scales, to establish connectivity among water polygons and flowlines and in so doing, to preserve horizontal data integration. At smaller mapping scales, a generalized primary path may substitute for the entire network. Stream order is one metric commonly utilized to delineate primary paths (Merwade et al 2005). Other prioritization methods that have been used for streams or river basins are the Pfafstetter system (Verdin 1997), watershed area (Ai et. al 2006), or upstream drainage area (Stanislawski 2009). Given the multi-scale condition of the HR NHD, UDA prominence estimates are being used for the HR NHD because UDA is normalized by area and better suited for flow networks with compilation inconsistencies. The experiment extends previous database enrichment to build cartographic centerlines (Anderson-Tarver et al 2011, 2012).

The primary objective of the research reported here is to generate an automatic workflow to delineate primary paths in production of a generalized version of 1M National Atlas hydrography, and assess the degree to which the set of primary paths generated from 24K data reflect existing National Atlas content and geometry.

A second question to be considered in this paper relates to linking data versions produced at source and target mapping scales. The obvious advantage relates to improving the efficiency and consistency of data updates. The development and adoption of Multiple Resolution Databases or MRDBs (Kilpeläinen 1997) builds upon Gruenreich's (1985) dual database architectures of Digital Landscape Models (DLM) and Digital Cartographic Models (DCM), intended to provide versions of vector data for mapping at multiple scales and for multiple purposes, with feature-level linkages maintained between versions (Sarjakoski 2007). One of the most compelling obstacles

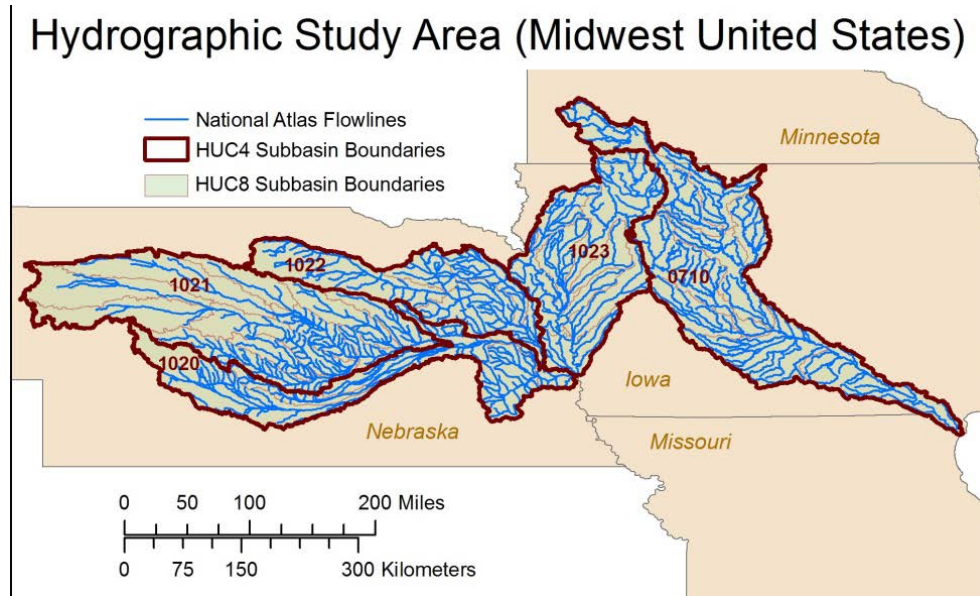
to a fully operational MRDB is the work involved to establish linkages among feature representations previously archived within isolated databases. Ideally, multiple representations should be linked during initial compilation, or at least the smaller scale versions should link back to larger scale versions. For National Mapping Agencies that maintain very large volume databases (terrabbytes or pedabytes of data), the task of combing through and associating individual features has obstructed production of a fully operational MRDB for many decades. By generalization from the largest scale hydrography available, a 1M version can be produced with feature-level linkages automatically carried from the 24K data into the generalized database.

## 2. Data Set

The NHD represents natural and human-made surface water features for the United States. Two versions are distributed by USGS: the HR NHD and the medium-resolution layer, which is compiled from 1:100,000-scale (100K) source data. A third version of hydrography has been derived from the 100K data and simplified to 1M for use in the USGS National Atlas® (Gary et al. 2010). The current experiment will generate a version of 1M data from the 24K source, to accomplish several objectives, namely to utilize a source dataset of higher positional accuracy than the 100K data. The 24K data is updated more frequently than the 100K, which will give the experimental 1M data improved currentness. A third advantage of using 24K source is to transfer permanent feature identifiers from the largest scale NHD to the smallest, providing feature level database linkages automatically to the National Atlas data.

The study area involves 36 hydrographic subbasins spanning a 140,172 km<sup>2</sup> largely agricultural region in the central United States (*Figure 1*). The landscape is generally flat or hilly but not mountainous, with precipitation runoff values ranging from 261 mm per year in the eastern tip (near the Iowa-Missouri border) to 19 mm per year in the drier western edge, in central Nebraska (Wolock and McCabe 1999). The average elevation rises steadily east to west, ranging from 200 meters above sea level to 1200 meters. Terrain is level in the northeast part of the study area, with an elevation standard deviation of roughly 2-5 m for 5km cells), rising to 13-33 m along the Missouri river bluffs between Iowa and Nebraska, and becoming varied in the central and western portions of Nebraska (7-23m standard deviation in elevation). Land use in the eastern half of this region is mainly machine-cultivated agriculture, and in the western parts it transitions to livestock

ranching. Elevation and standard deviation values are based on 5km averaging of 1:250,000 3-arcsecond DEMs.



**Figure 1.** A portion of the existing 1:1 million scale National Atlas hydrography, generalized from 100K NHD data is situated within thirty-six HUC8 subbasins and five HUC4 subregions, comprising the study area for this experiment.

The subbasins vary in stream channel density; in HUC4 #0710, larger scale hydrographic data show artifacts of a large glacial moraine, but this evidence is not apparent in the 1M data shown in *Figure 1*. Differences in drainage density are however apparent in three western subbasins, especially those within HUC4 #1021, and more subtle differences appear in subbasins within HUC4# 1023.

### 3. Generalization Processing

For the experiment, data will be enriched, ladder pruned and simplified to 100K. Primary paths will be delineated and then simplified to 1M. Resulting generalized channels will be conflated with existing National Atlas channels shown in *Figure 1*, to assess the generalization by comparing total stream length, average segment lengths, and establishing the extent to which these comparisons deviate across the study area.

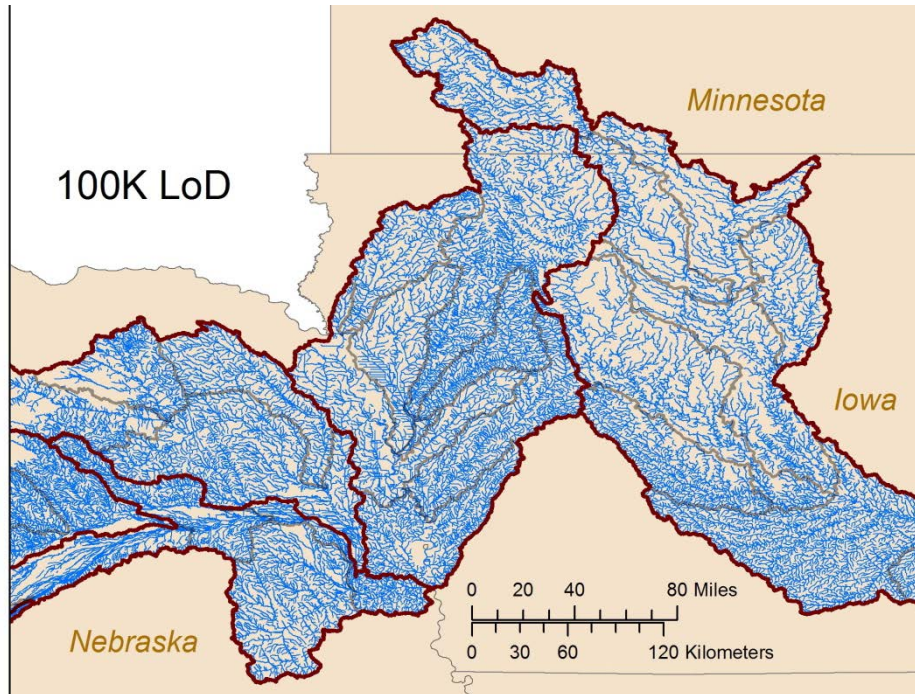
### 3.1. First Step on the Ladder: Generalization of the 100K LoD

The 24K data must be enriched prior to pruning and simplification. Database enrichment adds attributes to source scale data that support data characterization, local density estimation, algorithm and parameter choices, and vertical and horizontal data integration (Bobzien et al 2006, Steiniger and Weibel, 2007, Neun et al. 2008, Bittenfield et al., 2011). Enrichment for this experiment includes an estimate of upstream drainage area (UDA) for each stream channel, providing a relative prominence characterization for each feature. Others have used stream order or total upstream channel length (Thompson and Brooks 2000, Savino et al. 2011) to generalize stream channels, but these values are sensitive to inconsistent channel compilation, which does exist within the 24K NHD layer. UDA prominence estimates are normalized by area and are better suited for flow networks with compilation inconsistencies. Ai et al. (2006) also simplify a river network using watershed areas estimated through Delaunay triangulation. UDA estimates for the HR NHD are derived from Thiessen partitioning of catchments for each flowline feature (Stanislowski 2006).

Stream densities are calculated from the upstream drainage estimates, and stratified to guide adaptive pruning which maintains local density variations in the 100K LoD (Stanislowski and Bittenfield 2011). Partitioning has been used by other researchers to differentially process areas with different data densities (Bobzien et al. 2008, Chaudhry & Mackaness 2008, Stanislowski 2009) and to monitor isolated (point) objects over time and space (Downs 2010). Density strata were identified for the 36 subbasins, with upper limits averaging 0.0774, 0.7515, 1.3141, 4.5542 and 5.0843 km per sq km respectively. Most subbasins were characterized by stream densities for the lowest density categories (1,2,3).

Pruning involves elimination of confluence-to-confluence stream segments while preserving network connectivity. Pruning is stratified to different thresholds to preserve local density variations and maintain visual contrast in densities for the generalized data. Thresholds for pruning are guided by a variation in the Radical Law (Töpfer and Pillewizer 1966) which computes a change in stream channel density instead of a change in the number of stream channels. Radical Law values are computed for each density strata in each subbasin, and the values are adjusted by computing an “expansion factor” that compensates for the size of the scale jump (Bittenfield et al. 2010, 2011).

Flowlines were simplified to 100K using Wang and Muller’s (1998) Bend-Simplify algorithm with tolerance thresholds of 70m for the lowest density classes, and 50 m for all other strata (*Figure 2*). Flowlines in the 100K LoD display show that channel density differences are still apparent.



**Figure 2.** A portion of the 1:100,000 generalized NHD hydrography, generalized from HR NHD data. Differences in stream channel density are still apparent in this ladder step generalization. HUC4 and HUC8 boundaries are symbolized as in *Figure 1*.

In particular, one can see the toe of the glacial moraine in the eastern most HUC4. One can also visually distinguish between natural stream features and human-made features (e.g., irrigation canals and ditches) in the center of the display immediately west of the Nebraska-Iowa border. Water polygons were simplified in a two-stage process that eliminated smaller polygons on scale-dependent minimum area criteria developed specifically for NHD (USEPA and USDOI 1999), and simplified boundaries of remaining polygons. A check insured that waterbody polygons connecting stream channels were retained regardless of size, to preserve network continuity.

### 3.2. Second Step on the Ladder: Generalization of the 1M LoD

Reiterating from the problem statement above, the second ladder generalization will transform the 100K LoD to a level of detail appropriate to 1M. As

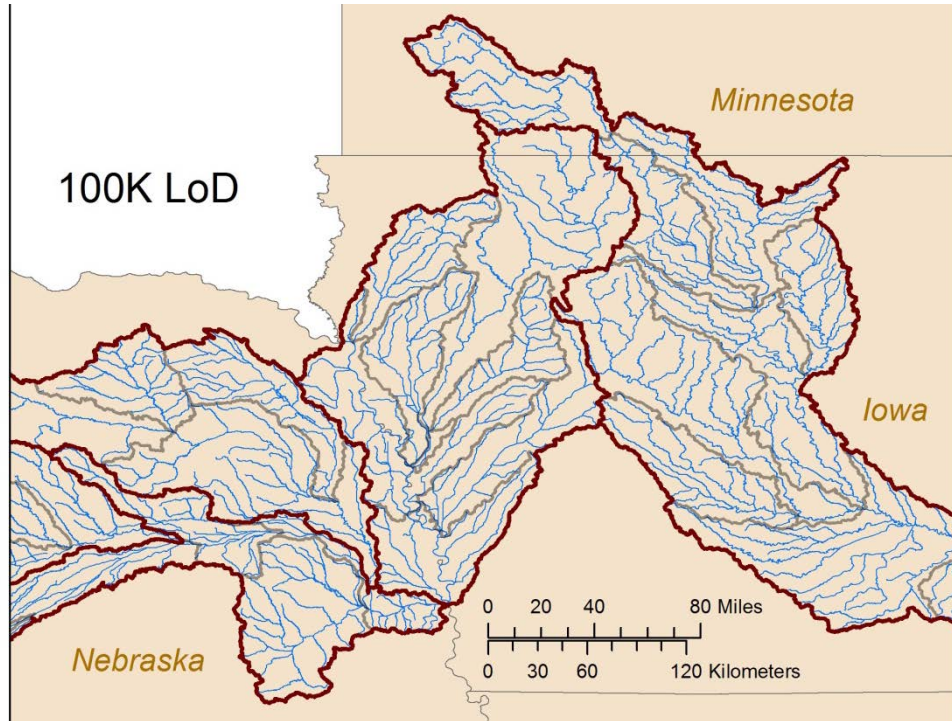
before, the processing will involve a sequence integrating pruning, elimination and simplification. Conventional pruning is not applied in creating the 1M LoD however because it tends to eliminate stream channel headwaters, which are considered significant components for a number of applications of National Atlas data. Instead, primary paths will be delineated using the 100K LoD which run from pour point to headwaters. The primary path network will be further processed to simplify and eliminate details, thus creating a 1M LoD.

The problem in accomplishing this step is that an attribute delineating primary paths is not incorporated into the HR data (and therefore not available in the 100K LoD), because of the expansive spatial footprint, which covers roughly 55,000 24K topographic map sheets for the coterminous United States, and because of the frequent and irregular update cycle, which for pragmatic reasons cannot be applied comprehensively to the national scale database. Primary paths cannot be derived from raster DEMs for this experiment, as the channels and reach codes must match HR NHD flowlines to link features in the two databases; thus primary paths must be delineated directly from the 100K LoD.

The primary path delineation first selects stream channels on the basis of the UDA attribute computed during enrichment. Channels which drain more than a specific percentage of a subbasin's total area are selected as primary path stems. For this experiment, a UDA threshold value of 1.5 percent was determined to generate a set of stems that most closely match the existing National Atlas stream network. The algorithm then runs a shared node trace upstream from the top of each UDA-selected stem to its headwaters, and this complete set of channels forms the network of primary paths. (*Figure 3*). Inspecting the figure, one will see that stream channel density differences are no longer apparent. These primary paths provide the pruning step in production of the 1M LoD.

Simplification of the primary path network to a level of detail appropriate for a 1M mapping scale involved analogous processing to the first ladder step, except that a single simplification tolerance of 500m was applied to all stream channels, since local density differences were no longer apparent. Waterbody polygons were also simplified to 500m tolerance, and smaller polygons were eliminated according to the same minimum area criteria described above.





**Figure 3.** Primary paths delineated from the 100K LoD. HUC4 and HUC8 boundaries are symbolized as in *Figure 1*.

#### 4. Validation

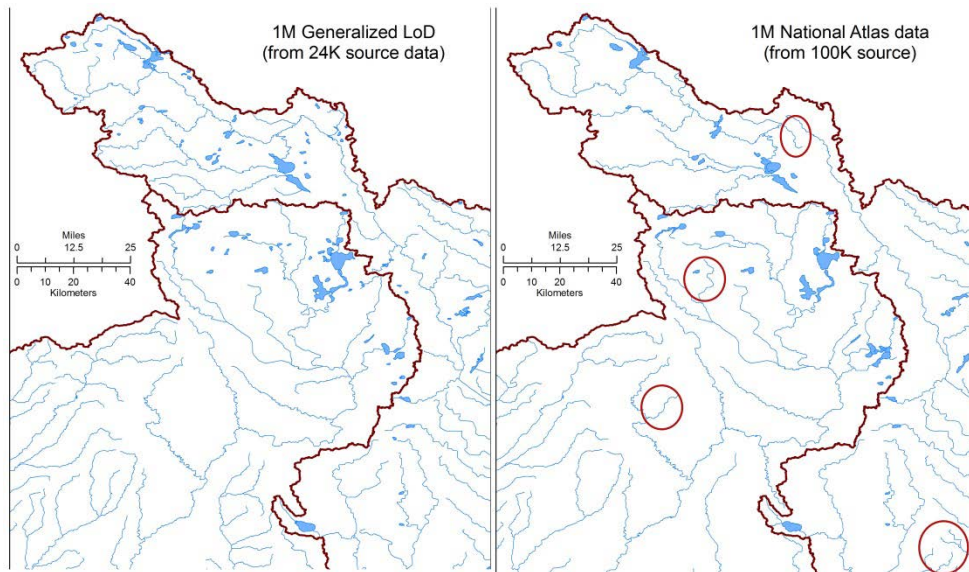
Visual comparison of the 1M generalized stream network with the National Atlas hydrography shows some differences in flowline channels and waterbodies (*Figure 4*). The 1M LoD was generalized from source data compiled at 24K, a scale four times larger than the source (100K) for the National Atlas data, so the additional channels and water polygons is not surprising. Channels which appear in the National Atlas data but not in the 1M LoD (examples circled in the figure) are characterized by low UDA values and were pruned during generalization of the 100K LoD, prior to delineation of a primary path.

The 1M generalized stream network was also compared with the existing 1M National Atlas data by means of conflation analysis. A conflation metric called the Coefficient of Line Correspondence (CLC) was adapted by Stanislawski et al. (2010) from a measure for areal feature coincidence discussed by Taylor (1977). CLC measures the proportion of flowline summed lengths



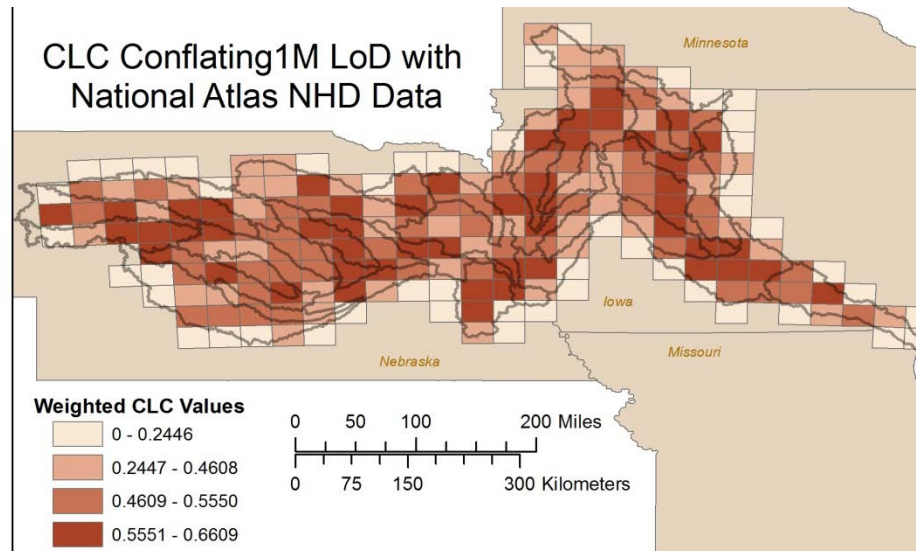
which match in the LoD and the benchmark with respect to the summed length of all channels. Perfect correspondence gives a CLC of 1.0; and a total mismatch between the two data sets is indicated by a CLC of 0.0. The CLC value aggregated over the 36 subbasins is 0.7760, indicating that more than  $\frac{3}{4}$  of the confluence to confluence stream channels in the 1M LoD match corresponding channel positions in the National Atlas.

**Figure 4.** A portion of the 1M data generalized from 24K source (on the left) com-



pared with the National Atlas data generalized from 100K source (on the right), displayed at 1:1,000,000. The portion shown lies in the northern part of two HUC4's (#0710 and #1023, labeled in *Figure 1*). Red circles indicate locations where the National Atlas incorporates channels not evident in the 1M LoD.

To explore the spatial pattern of line feature correspondence, a grid of 200 cells was overlaid over the study area to sample line features and compute a CLC value for each cell. Grid cell CLC values are weighted by the amount of area in each cell which contains any line features at all. The aggregate weighted CLC totals 0.7899 for the entire study area; and the gridded values are illustrated in *Figure 5*. Extremely low CLC values are evident around the edges of the study area, and low values can also be seen in the western drier part of the study area, where hydrographic channels become quite sparse.



**Figure 5.** Weighted CLC values for grid cells covering the 36 subbasins that conflate the 1M LoD primary paths with the existing National Atlas stream channels illustrated in *Figure 4*. CLC values are expressed as percentages, i.e., the highest value in the legend reflects a weighted CLC of 0.6609%, relative to the amount of area in each grid cell covered by stream channels.

## 5. Summary

This paper explains the processing mechanics for producing a 1M LoD of NHD vector data from HR source data, for use in the USGS National Atlas. Three advantages are gained by using HR data as source. First, the generalized data carries improved positional accuracy compared to existing National Atlas hydrography, because the HR data is compiled to a more precise standard than the 100K data which serves as source data for the current version. Second, the 1M LoD carries identical feature identifiers, which link the 1M LoD with HR source at the feature level. Feature level linkages provide an important prerequisite to production of a fully operational MRDB, which has proven to be a challenge for USGS data production in past years.

## Acknowledgements

The research by the Colorado researchers is supported by USGS-CEGIS grant # 04121HS029, "Generalization and Data Modeling for New Generation Topographic Mapping".

## References

- Ai T, Liu Y, Chen J (2006) The Hierarchical Watershed Partitioning and Data Simplification of River Network. In: *Progress in Spatial Data Handling Part 11*, p.617-632, doi:10.1007/3-540-35589-8\_39
- Anderson\_Tarver, C, Gleason, MJ, Buttenfield, BP, Stanislawski, LV (2012) Automated Centerline Delineation to Enrich the National Hydrography Dataset. *Proceedings GIScience 2012* Columbus, Ohio: Springer LNCS 748:15-28
- Anderson-Tarver, C and Buttenfield, BP Stanislawski, LV Koontz, J (2011) Automated Delineation of Stream Centerlines for the USGS National Hydrography Dataset. *Proceedings 25<sup>th</sup> International Cartographic Congress* Paris France, Vol 1:409-423
- Bobzien, M, Burghardt, D, Neun, M, Weibel, R. (2006) Multi-Representation Databases With Explicitly Modeled Horizontal, Vertical and Update Relations. In: *Proceedings AutoCarto*, Vancouver (2006)
- Buttenfield, BP, Stanislawski, LV and Brewer CA. (2011) A Comparison of Star and Ladder Generalization Strategies for Intermediate Scale Processing of USGS National Hydrography Data Set. *Proceedings, 14th Workshop ICA Commission on Generalisation and Multiple Representation*, Paris France, July
- Buttenfield, B.P., Stanislawski, L.V., Brewer, C.A. 2010. Multiscale Representations of Water: Tailoring Generalization Sequences to Specific Physiographic Regimes. *GIScience 2010*, Zurich, Switzerland, June
- Chaudhry, O and Mackaness, WA. (2008) Partitioning to Make Manageable the Generalization of National Spatial Datasets, *Proceedings 11<sup>th</sup> ICA Workshop on Generalization and Multiple Representation*, Montpellier, France.
- Downs, J. 2010 Time-Geographic Density Estimation for Moving Point Objects. *Proceedings of the 6th International Conference on Geographic Information Science (GIScience 2010)*, Zurich, Switzerland
- Gary R H, Wilson Z D, Archuleta C-A M, Thompson F E, and Vrabel J (2010) Production of a National 1:1,000,000-Scale Hydrography Dataset for the United States—Feature Selection, Simplification, and Refinement. *Scientific Investigation Report 2009-5202*, U.S. Department of Interior, U.S. Geological Survey
- Gruenreich D (1985) Computer-Assisted Generalization. In: *Papers CERCO-Cartography Course*. Frankfurt am Main, Germany: Institut für Angewandte Geodäsie (renamed Bundesamt für Kartografie und Geodäsie in 1997), 19 pp
- Kilpeläinen, LT (1997) *Multiple Representation and Generalization of Geodatabases for Topographic Maps*. Helsinki, Finland: Finnish Geodetic Institute
- Merwade, VM, Maidment, DR, Hodges, B. (2005) Geospatial Representation of River Channels. *Journal of Hydrologic Engineering* 10(3):243-251
- Neun, M, Burghardt, D, Weibel, R. (2008) Web Service Approaches for Providing Enriched Data Structures to Generalisation Operators. *International Journal of Geographical Information Science* 22(2):133-165

- Sarjakoski LT. (2007) Conceptual Models of Generalisation and Multiple Representation., In Mackaness WA, Ruas A, Sarjakoski LT (eds), *Generalisation of Geographic Information: Cartographic Modelling and Applications*, Elsevier: 11-35
- Savino S, Rumor M, Canton F, Langiu G, Reineri M. (2011) Model Generalization of the Hydrography Network in the CARGEN Project. In: Ruas A (ed), *Advances in Cartography and GIScience* (1):439-457, Lecture Notes in Geoinformation and Cartography. Berlin: Springer-Verlag.
- Stanislawski, LV. (2009) Feature Pruning by Upstream Drainage Area to Support Automated Generalization of the United States National Hydrography Dataset. *Computers, Environment and Urban Systems*, 33(5): 325-333.
- Stanislawski, LV and Battenfield BP. (2011) A Raster Alternative for Partitioning Line Densities to Support Automated Cartographic Generalization. *Proceedings 25<sup>th</sup> International Cartographic Congress* Paris France.
- Stanislawski, LV, Battenfield, BP and Samaranayake, VA. (2010) Automated Metric Assessment of Hydrographic Feature Generalization Through Bootstrapping. *Proceedings, 13th Workshop of the International Cartographic Association Commission on Generalisation and Multiple Representation*, Zurich Switzerland. [http://ica.ign.fr/2010\\_Zurich/genemr2010\\_submission\\_11.pdf](http://ica.ign.fr/2010_Zurich/genemr2010_submission_11.pdf)
- Steiniger, S, Weibel R. (2007) Relations Among Map Objects in Cartographic Generalization. *Cartography and Geographic Information Science* 34(3), 175-197
- Taylor, PJ. (1977) *Quantitative Methods in Geography: An Introduction to Spatial Analysis*, Chapter 5: Areal Association. Boston: Houghton Mifflin
- Thomson, RC, Brooks, R. (2000) Efficient Generalization and Abstraction of Network Data Using Perceptual Grouping. In: *Proceedings of the 5th International Conference on GeoComputation*
- Töpfer, F, Pillewizer, W. (1966) The Principles of Selection. *The Cartographic Journal* 3: 10-16
- USEPA and USDOI. (1999) Standards for National Hydrography Dataset, United States Environmental Protection Agency and United States Department of the Interior, United States Geological Survey, National Mapping Program Technical Instructions, July. <http://rockyweb.cr.usgs.gov/nmpstds/acrodcs/draft/dlg-f/nhd/NHD0799.pdf>
- Verdin, KL. (1997) A System for Topologically Coding Global Drainage Basins and Stream Networks. In 1997 ESRI international GIS user conference proceedings. <http://gis.esri.com/library/userconf/proc97/proc97/to350/pap311/p311.htm>
- Wang, Z and Muller JC. (1998) Line Generalization Based on Analysis of Shape Characteristics. *Cartography and Geographic Information Science* 25(1): 3-15
- Wolock, DM and McCabe, GJ. (1999) Estimates of Runoff using Water-balance and Atmospheric General Circulation Models. *Journal of the American Water Resources Association*, 35(6):1341-1350